Team Progress Report for November 19th, 2023

Original Project Proposal:

- Team PST Powerhouse Project Proposal.pdf

Team Members:

- elm3
- arjun12
- ml171
- yongy3
- ganesan8

Current Project Status:

Presently, the team is meeting twice a week for live discussion and task distribution. Progress has been steady, with the following portions either done or soon to be merged:

- Dataset selection
- Dataset ingestion and cleanup
    - Implemented with the Python Natural Language Toolkit library.
    - Includes removing stop words, stemming, removing symbols, etc.
- Information Retrieval and Ranking
    - Implemented with the rank-bm25 library and using the Okapi Bm25 algorithm.
    - Initial testing is promising, with reviews for each title being ranked in aggregate for a given query.
- Sentiment analysis
    - Mostly finished, using the Flair NLP Library for analyzing the positivity of game reviews.
    - Accuracy for sentiment analysis is currently over 80%.
    - For efficiency, we will be moving to sampling reviews for sentiment analysis, rather than the entire corpus of documents.
- Frontend
    - Initially built two prototypes in python and JavaScript, with the team ultimately deciding to use JavaScript.
    - Aim to finish the front-end interface by next week.

The following tasks remain:

- Completing the user interface.
- Integrating the disparate components together (data cleaning, ranking, sentiment analysis, and frontend)
- Full integration testing for the application.
- Automating testing of precision and recall, sentiment analysis.
- Preparing presentation materials.

Project timeline:

Week of 10/09:

- Team introductions
- Brainstormed project topics, voted on current topic and formulated a proposal.

Week of 10/23:

- Started investigation of appropriate data sets.
- Started investigation into algorithms and backend libraries.
- Started investigation of frontend libraries.

Week of 10/ 30:

- Decided on using the current Kaggle dataset for steam reviews.
- Decided on using Okapi BM25 for query ranking, Flair NLP for sentiment analysis.
- Split prototyping for the frontend between Tkinter for Python and JavaScript.

Week of 11/06:

- Implemented data ingestion/cleaning for the dataset.
- Started implementing sentiment analysis, using sample datasets to test accuracy.
- Stared implementing code for Okapi BM25 for use with our dataset.
- Continued frontend prototyping.

Week of 11/13:

- Finished initial code for sentiment analysis and retrieval/ranking with BM25.
- Demoed frontend prototypes, deciding to go with JavaScript.
- Added improvements to data cleaning, removing 'Early Access' reviews that did not contain usable text information.
- Started integration of data cleaning, IR and ranking, and sentiment analysis (early testing).